DOCUMENT RESUME

ED 162 624 IR 006 341

AUTHOR Scholtz, Sandra
TITLE Data Structure and Computerized Museum Catalogs. Museum Data Bank Research Report No. 2.
INSTITUTION Margaret Woodbury Strong Museum, Rochester, N.Y. Museum Data Bank Committee.
PUB DATE Nov 74
NOTE 10p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Cataloging; Comparative Analysis; *Electronic Data Processing; Methods; *Museums; *Structural Analysis; Structural Linguistics

ABSTRACT

The nature of the structure in museum information is explored, and some alternative techniques for maintaining this structure in computerized records are compared. One approach uses structural linguistic concepts to describe relationships in the information from which museum catalog data are extracted and expands these data to an appropriate host language statement. This expanded expression, the full terminal string, can be paraphrased but it cannot be further expanded to change its meaning. The full terminal string from which a catalog record is extracted is hierarchical with various positions called nodes. Relationships between terms in a full terminal string may be relationships of co-occurence--simple dependence, mutual dependence, or mutual exclusion; conjoining; or embedding. These relationships must be retained when extracting catalog data in order to accurately represent the original data meanings. Two approaches for handling these structural relationships when extracting data are compared; one keeps context and structure separate from the extracted catalog data in data category definitions, while the other takes advantage of implied context data, such as general record subjects, and extracts other contexts as part of the data. (Author/MBR)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

- ED162624

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

MUSEUM DATA BANK RESEARCH REPORT

Number 2

Data Structure
and Computerized Museum Catalogs

by

Sandra Scholtz

November 1974

This report, or any portion thereof, may be reproduced for any non-commercial purpose provided appropriate credit is given to the author and the Museum Data Bank Committee as publisher.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. G. Chenhall

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Museum Data Bank Committee
Robert G. Chenhall, Chairman

Margaret Woodbury Strong Museum
700 Allen Creek Road
Rochester, N.Y.

IR001341

Alfred North Whitehead once said that all of the important things had been said by people who did not discover them. This paper will begin by discussing the application of some structural concepts to computerized museum catalog records. These concepts were not discovered by me, but I believe them to be of sufficient importance to warrant saying at this time. I have found that these concepts have given me an integrated terminology for describing structural relationships between data items in a catalog record, a terminology for which I had felt a lack before. Following the definitions of the structural concepts will be a comparison of two approaches to handling these relationships in computerized museum catalogs.

The information that is recorded on museum catalog cards and which may be converted to machine readable form constitutes a language of communication which is designed and maintained by appropriate museum functionaries and which must be understood by those who desire to use the catalogs. The cataloging language is simply a derivative of the host language itself, which in our case is English. Vance (1973: 27-28) has demonstrated the expansion of museum catalog data into an adequate English expression for clarifying the derived catalog data. His example is:

ARTIST: LA FRESNAYE, Roger de (French, 1885-1925).

which is the cataloged data for the item registered as number 222.47. He expands this into the English phrase which follows:

> Number 222.47 is a work of art, which in the considered opinion of its present cataloger was conceived and executed by M. Roger de La Fresnaye, who by the way was a citizen of France, born in the year 1885 of the Christian Era and deceased in the year 1925 of the same Era.

These same facts could, of course, be adequately clarified by any number of paraphrased or transformed English expressions and this is immaterial as long as the meanings of the various data items are maintained accurately. The point that is being stressed is that when one learns to use a particular museum catalog he is actually internalizing the rules for expanding derived catalog data into an appropriate English expression.

Examples of much more complicated catalog information could be given readily. The information about an item with separate describable parts, such as an Eskimo needle case with its contents, could easily extend over several catalog cards and would require rules for maintaining in the derived data the structure of the information in the English expression. Without the proper maintenance of this structure the derived data would become garbled and an accurate interpretation of the information would be impossible.

The people who have designed museum cataloging systems have devised techniques for maintaining data structure in their derived catalog data. It is just as necessary for those of us who are involved with converting this catalog data to machine readable form to devise techniques for maintaining data structure in the computerized record. The intent of this

paper is to explore in some depth the nature of the structure in museum information and to compare some alternative techniques which have been devised for maintaining this structure in computerized records of the information.

The position to be presented here is that the structure in museum information is nothing more than a reflection of the structure of this information expanded in the host language. The structural concepts that will be described below for expressing the various relationships between elements of a museum catalog record are the same concepts that structural linguists some time ago applied to the analysis of the structure in language (e.g., Koutsoudas 1966). These concepts need to be defined and their application to museum catalog information clarified.

It has been stated above that the data in a museum catalog record can be expanded to an appropriate English language statement. The linguistic term for this expanded expression is terminal string, that is, the expression can be paraphrased or transformed without changing meaning, but it cannot be further expanded to add to the meaning. In most museum situations the terminal string is probably never written in full form; nevertheless,it exists in the mind of the person who extracts the data that are recorded. In English the complete terminal string with which we deal is a sentence. From this point of view a definition of a museum catalog record is an extraction of appropriate items of data with a maintenance of structure from a terminal string or sentence. This sentence may contain information about one or many different objects or places, for example, and may range from a very concise expression conveying only a single fact to an extremely complicated expression conveying many facts. Each extracted item of information could be expanded into a clause; one such clause can constitute a sentence or record. The data which are considered appropriate for inclusion in a museum catalog are determined by the purposes which the catalog is intended to serve.

The full terminal string or sentence from which a catalog record is extracted is hierarchical or tree-structured, as is any other sentence. In a linguistic hierarchy the various possible positions are called nodes. The nodes which occur toward the top of the tree are referred to as higher order constituents, while those nodes occurring toward the bottom of a tree are referred to as lower order constituents. A node consists of one or more basic linguistic units (morphemes), or one or more items of data in a catalog record.

The terms defining the structural relationships in a sentence are appropriate for defining the structure in a catalog record. These relationships are referred to as the relationships of co-occurrence, those of conjoining, and those of embedding.

Relationships of co-occurrence are those which relate to the dependence between linguistic morphemes or between data items in a catalog record. Three different types of co-occurrence relationships are recognized by linguists. The first type of co-occurrence relationship is a

simple (or unidirectional) dependence between two or more units. An example from the cataloging of a prehistoric collection involves the dependence of a unit specifying techniques of manufacture upon the unit which identifies the object being cataloged. If the object is identified as pottery then it is appropriate to observe a technique of manufacture such as coiling. Other units of observation may become appropriate also, such as surface finish or kind and amount of temper. If however the object being cataloged is identified as an unmodified fragment of deer antler, then there is no technique of manufacture, surface finish, or temper. The relationships between the unit, pottery, and the unit, coiling, is one of simple dependence, the presence of the latter dependent upon the occurrence of the former.

The second type of co-occurrence relation is a mutual (or bi-directional) dependence between units. An example is the observation of a unit specifying name and one specifying address when recording a contact who can supply additional information about an archeological site. Either of these units alone would not meaningfully identify a contact, so their relationship is one of mutual dependence. Another example of mutual dependence between units is seen to exist between a unit specifying latitude and one specifying longitude in giving a location.

The third type of co-occurrence relation is a mutual exclusion between units such as a unit specifying a date of manufacture of a man-made object and a unit specifying the stratigraphic association of a fossil. These two units would never appear in the same record as long as the man-made object and the fossil are cataloged separately.

It should be observed that co-occurrence relationships between units are defined within a particular context in which the units occur. For example, in the context of archeological site records, a person's name may co-occur with his mailing address--the name and address of the site reporter, the recorder, the owner, and the tenant. In the context of a work of art, however, the name of the artist will co-occur with his country of origin.

The concept of co-occurrence relationships between data units appears in Cutbill's work as the "group" concept (Cutbill n.d.). A "statement group" consists of associated or linked statements or units of data within a particular group context. The group contexts involve events of collection, of cataloging or of identification of the object.

The relationships of conjoining between units of data in a catalog record are those combinations of units which are the result of combining similar clauses in the same record. For example there may be numerous kinds of features found at any one particular archeological site--mound, fire pit, post mold, and house pattern to name but a few. These units of data may be considered conjoined units in the catalog record for the site in which they occur. Other examples are multiple

artists for a single work of art, multiple contacts for a single archeological site, or multiple owners for a historic landmark. Conjoined elements occupy similar positions in an hierarchical structure.

The final relationship to be defined is that of embedding. In linguistic terminology the embedded clause of a complex sentence is referred to as the dependent or subordinate clause. In catalog records, units of data may also be embedded. An example is the observation that an archeological site which is the subject of a catalog record is characterized by a kind of feature called a mound which has dimensions of its own, distinctive from the dimensions of the site. The unit specifying the dimensions of the mound is related by embedding to the unit identifying a mound as a feature present. Embedding relationships are represented linearly in a hierarchical structure.

In a museum catalog record there are likely to be a number of different units of data which are associated in a structure by relations of co-occurrence, conjoining and embedding. The position that has been taken here is that the structure of the catalog record reflects the structure of the full terminal string or sentence from which it was derived. A point which needs to be stressed is the importance of context in determining this structure. A unit of data which is related to another unit by simple dependence, must become embedded in the structure if the unit to which it is related is conjoined with other similar units, perhaps with their own dependent data. The advantages of the implications of context are valuable and should be exploited. Probably the most obvious use of context is that of the subject of the record itself. It is probably fairly standard to divide museum catalogs into separate computerized files based on similarity of record subject. At the University of Arkansas Museum we either have or are planning a separate file for archeological site information, one for photographic slides, one for black and white photographs, several for the biological collections based on taxonomic division at the class level, one for the geological collection, and ultimately will surely have many separate files for the massive archeological collection. Units of data in each of these files occur within the context of the general subject matter of the file. Context within the structure of the record is of equal importance, and this discussion will come back to this point.

With respect to the preceding discussion, I would define museum catalog data as nothing more than units or bits of information extracted from terminal strings of pertinence to a museum catalog. The data are the units of information themselves, and the information itself should be carefully distinguished from the set of rules by which the catalog data are extracted. There is no one inherently natural or logical set of rules for extracting catalog data, but any set of rules designed for this purpose must allow the user to retain in the extracted data the structure which is present in the terminal string. Various techniques have been devised for maintaining this structure in card catalogs, such as headings on the cards, positional relationships on the cards, the use

of additional cards for describing parts, and other such techniques. In converting these records to machine readable form, effective techniques must be devised for preserving the structure of relationships between data items. In other words, there must be a set of rules by which one is to extract the catalog data while maintaining the data structure. Two approaches for extracting catalog data and maintaining data structure which have been applied in computerized museum catalog records will be compared. They differ mainly in the techniques which have been devised for handling and using context.

One possible approach is that in which computerized records are structured in such a way that context is treated as part of the extracted data and appears in the record in the form of keywords to which the relationship of other units is defined by their position in the structure relative to these context keywords. One application of such an approach is the recording format designed by Cutbill (n.d.; 1973) in which there is a standard format for recording common data units or sets of co-occurring units along with keywords which indicate context or position of the units in the data structure. There is no data category definition list to grow; instead there is a context keyword list which can expand and can be handled as any lexicon. Instead of a definition for every possible kind of date, for example, there is simply a standardized method for recording a date, and the meaning of a particular date is interpreted by its relationship to a context keyword.

To be contrasted with this is the approach in which numerically tagged categories of data are defined and the structural relations between data elements and the contexts in which they are to be interpreted are part of the computerized records. An application of this approach is that of the Museum Computer Network (Vance 1973; 1974). The members of MCN follow a set of rules provided by the Network for creating computerized catalogs with the program system called General Retrieval and Information Processor for Humanities Oriented Studies or GRIPHOS (Heller 1974). The MCN data category list is becoming quite extensive and surely will continue to grow because each slightly different context requires that a new category be defined. For example, the recording of names and corresponding addresses in an archeological site file requires quite a few different tag numbers depending on the context of occurrence. The contexts in which names and addresses are used include that of reporter, recorder, owner, tenant; other possible contexts which have been suggested are those of informant, contact, excavator, and collector.

In addition to becoming extensive, the MCN data category list is complicated in that it is divided into several series essentially for different record subject contexts. What this means is that the same item of data may be defined in each of the series, and thus have a different tag number for each different context. For example, a different tag number is required for recording an archeological site number in the context of a site record and in the context of an object record. If one wanted to record a site number in the context of a

biographical record still another definition and tag number would be necessary. None of these tag numbers for site numbers could be used in a record concerning a document or a concept. At present there are six different series, one for each of six kinds of entities, which are object, person or group, document, place, event, and concept (Vance 1974: 3). My own interpretation of these six entities is that they are simply six possible record subjects, or units of observation. They represent one of many possible ways of dividing up the whole of reality, into which everything does not comfortably fit. Thus problems of pigeon holing are created such as where to fit a building into this division, as a place or an object. I am of the opinion that many units of observation, or record subjects, will be defined and this list of series logically would have to continue to grow. Such a unit of observation as a stratigraphic unit within an archeological site, for example, does not fit into this list of six entities. The stratigraphic unit would be the organizational unit for data concerning characteristics and content of the soil, observations as to any features such as pits or post molds, as well as observations as to the kind and amount of various artifact types found within the unit. Would this require a new series of tag numbers? Would the unit "site number" need yet another tag since the context in which it is recorded is slightly different from that in any of the existing definitions? Possible units of observation, or record subjects, are as numerous as possible research designs. The maintenance of separate series of tag numbers for each possible record subject cannot help becoming immeasurably complicated for museums in which scientific research will be the reason for much of the computerization of information concerning the collections.

The GRIPHOS program package allows for the creation of structured records in an extremely flexible manner. Records can be structured to represent any relations of co-occurrence, conjoining and embedding between data elements, and there is no restriction requiring symmetry in the structure. The added feature of threading, or connecting records in separate files, allows one to treat entire records in one file as embedded data in appropriate records in a file to which they are threaded. In addition, there is a feature of structured GRIPHOS records which could be designed to serve a number of different purposes --hierarchical node names are themselves a searchable category.

I would like to suggest the desirability of combining a program package, such as GRIPHOS, with its potential for handling complex structures, with a method of catalog data extraction in which context was a part of the data itself. This approach would allow one to partition or subdivide a record into different kinds of information, such as categories of data concerning the cataloging of the item, those concerning the history of the item, those of physical description and identification, those concerning conservation information, and any other logically co-occurring groups of data categories. Such partitioning could be designed to allow one to process parts of records only and this would make computerized catalogs more efficient to use if the

partitions were planned to reflect the information needs of the institution creating the computerized catalog. This approach also would allow a virtual elimination of the data category definition lists and thus of the problems that come with them. To show an example using GRIPHOS, it is suggested that hierarchical node names could represent context keywords. Instead of numerical tags, words could be used for tags--words which indicate the nature of the unit of data. The tag DATE could serve for any date and the context in which this date is to be interpreted would be indicated by a node name such as CATALOG which would identify and group all co-occurring data items having to do with the cataloging operation. The cataloger's name could co-occur with the date and have the tag NAME. So instead of having several different tag numbers for dates with different contexts, and probably duplicates in several other series for some of those, there could be only one tag for a unit which specified a date--DATE. An index of node names could allow the location of the part or parts of records of pertinence to each instance of processing the data in the file. For example, one could design a partition of conservation data and produce periodic conservation reports without processing the remainder of the computerized records. The ability to create such a structure exists in GRIPHOS, however the retrieval system has not been designed to be used in this manner.

Logically the example that I give is with reference to the GRIPHOS system since it is the one with which I am most familiar. However, the question that I am raising is not concerned specifically with how GRIPHOS may or may not be used, but rather with the desirability of putting more structure into computerized catalog records. There is a need to re-examine our rules for catalog data extraction. The MCN techniques of data extraction were designed for art museums with a keen interest in networking capabilities, a subject discussed in some detail by Chenhall (1974), and I am sure that the design is well suited to the purpose. However, the possibility should be considered that a different design might better serve the needs of the general museum with extensive and highly varied collections and with interests in internal control of collections for more efficient use of these collections for exhibit, teaching, and research.

In this paper I have attempted to demonstrate that some structural linguistic concepts are useful for describing the relationships in the information from which museum catalog data are extracted. I have attempted to stress the point that these relationships must be retained when extracting catalog data in order to represent accurately the original meanings of the data. Two approaches to handling these structural relationships when extracting data for computerized catalogs have been described. I hope to elicit discussion on the applicability of these linguistic concepts to museum catalog data, as well as on the relative merits of the described alternative approaches to handling structural relationships: that of keeping context and structure separate from the extracted catalog data, in data category definitions, as opposed to taking advantage of implied context such as general record subject and extracting other contexts as part of the data.

REFERENCES CITED

Chenhall, Robert G.

1974 Museum information networks. Museum Data Bank Research Report 4, Museum Data Bank Committee, Rochester, N.Y.

Cutbill, John L.

1973 Computer Filing Systems for Museums and Research. Sedgwick Museum, Cambridge, England.

n.d. Untitled manuscript on computer filing systems, in preparation for publication, 43 pp.

Heller, Jack

1974 On logical data organization, card catalogs, and the GRIPHOS management information system. Museum Data Bank Research Report 3, Museum Data Bank Committee, Rochester, N.Y.

Koutsoudas, Andreas

1966 Writing Transformational Grammars: An Introduction. McGraw-Hill Book Company, New York.

Vance, David

1973 Computers in the Museum. IBM Data Processing Application No. GE20-0406-0. International Business Machines Corp., New York.

1974 Manual for MCN Data Preparation. Computer prepared manuscript, 27 pp.